Mateusz Wasiluk

ROBIN Lab Meeting
09.02.2023

Bayesian Neural Networks for Continual Learning

Plan for today

- The principles of Bayesian learning
- Inference methods for BNNs
- Mean field approximation in practice
- Introduction to normalizing flows

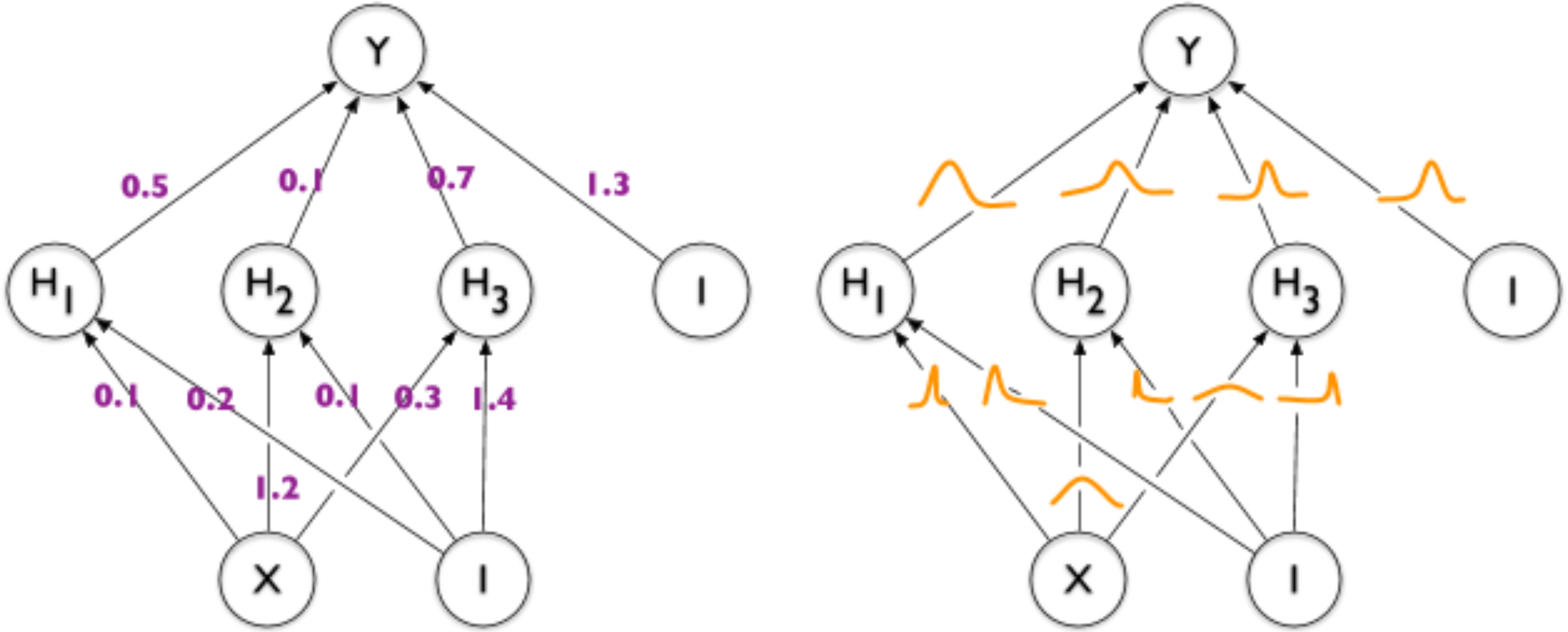
New observations influence our model of the world

$$Posterior \propto prior * likelihood$$

The new model depends on the old beliefs and the new observations *in the light of the old beliefs*

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} = \frac{P(D,H)}{\int_H P(D,H')dH'}$$

A BNN is any stochastic neural network trained with Bayesian Inference



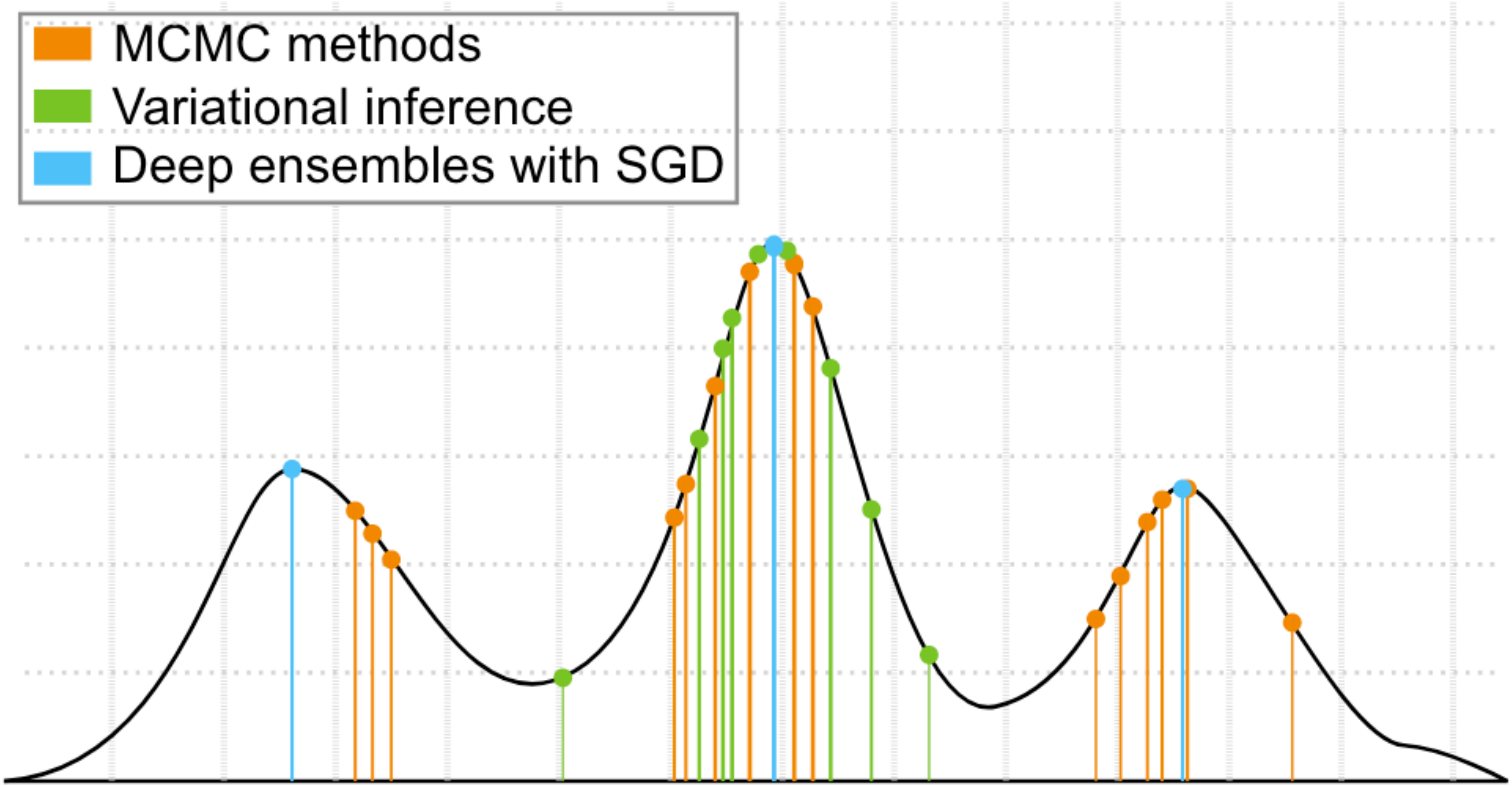
Most popular methods of inference are MCMC and VI

Markov Chain Monte Carlo

- Directly samples the posterior
- Can explore different modes
- Does not scale well

Variational Inference

- Uses a parametric approximation of the posterior
- Aims at minimizing the KL-divergence
- Scales for large models with a limited expressive power

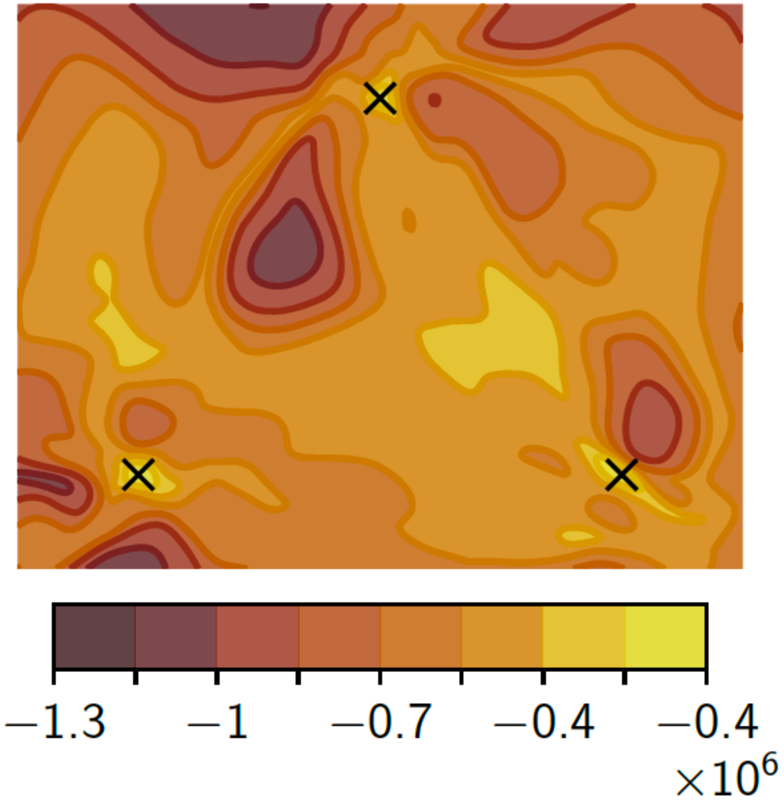

MCMC methods
Variational inference
Deep ensembles with SGD

Training a variational network requires maximizing the ELBO

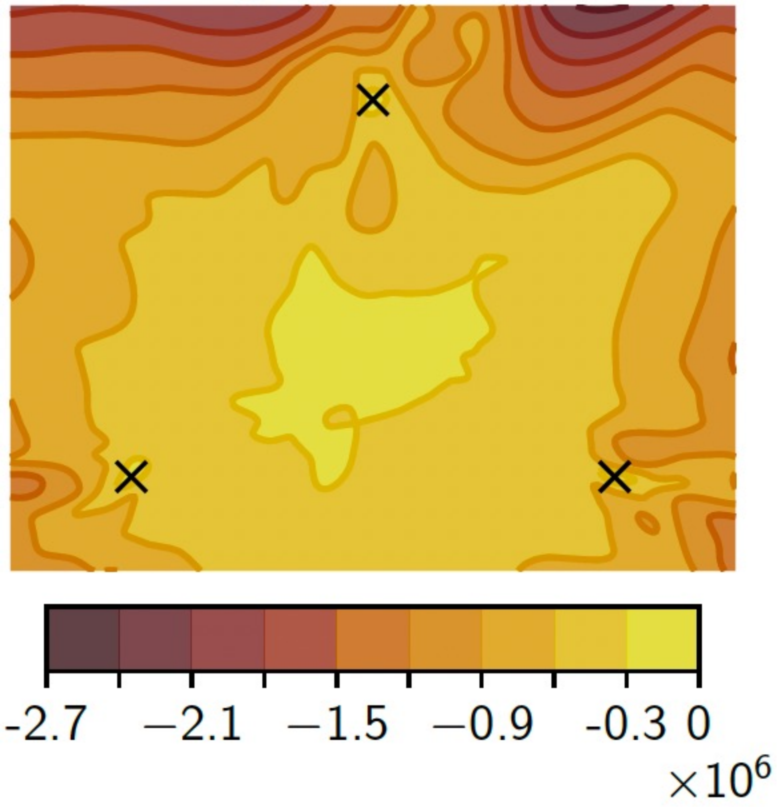
$$\underbrace{\overbrace{\log P(D)}^{\text{evidence}} - \overbrace{D_{KL}[Q(\theta)||P(\theta|D)]}^{\text{variational-posterior divergence}}}_{\text{ELBO (training loss)}} = \underbrace{E_{\theta\sim Q}[\log P(Y|X,\theta)]}_{\text{classification loss}} - \underbrace{D_{KL}[Q(\theta)|P(\theta)]}_{\text{variational-prior divergence}}$$

Mean-field assumption is too strong for complex datasets

Log-posterior, CIFAR-10, samples from one HMC chain

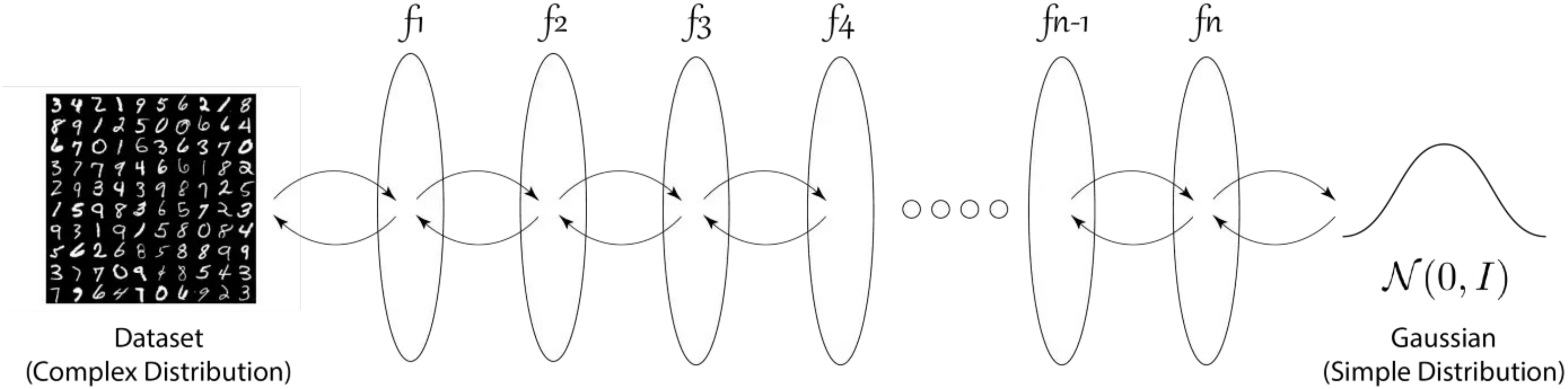


Log-posterior, CIFAR-10, samples from independent HMC chains

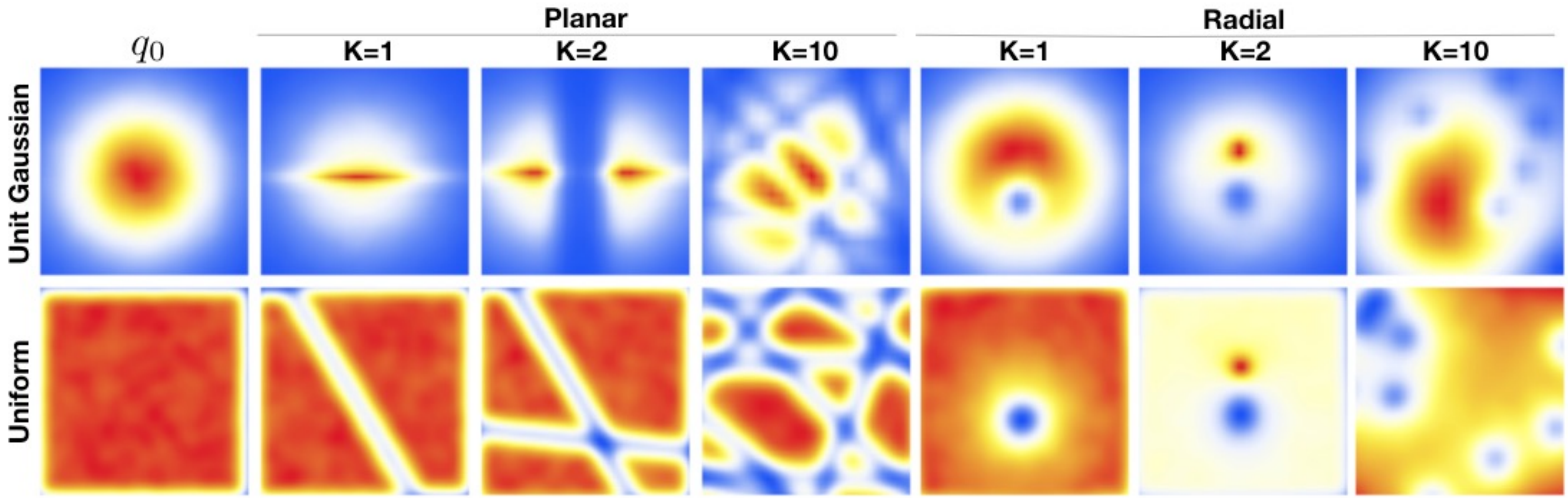


Figures adapted from Izmailov2021

Normalizing flows transform simple distributions to more complex ones



Example: ”geometrical” flows



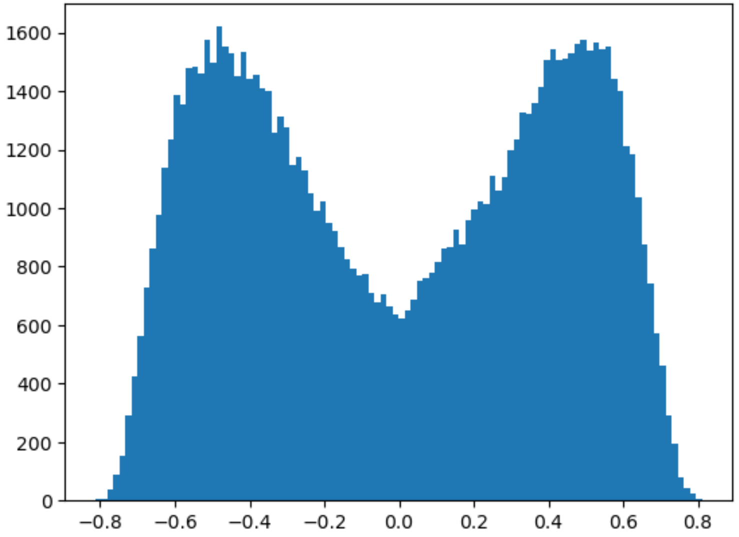
Figures adapted from Rezende2015

Want to understand more?

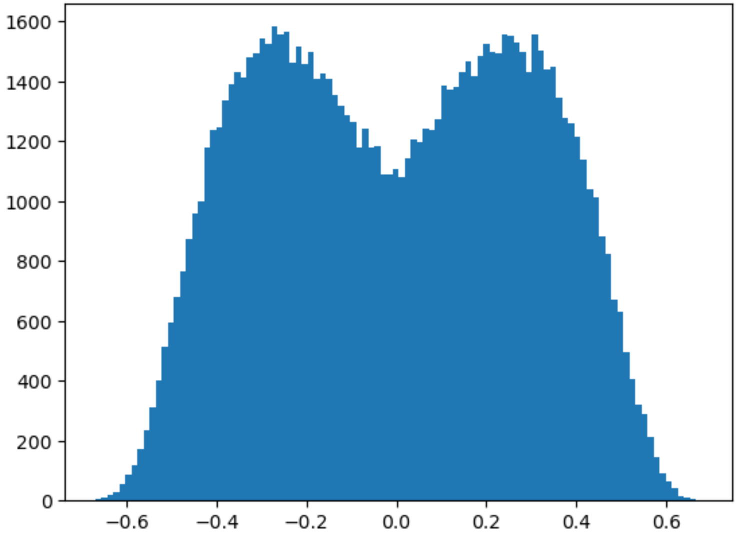
- Rezende, D., Mohamed, S. (2015). Variational Inference with Normalizing Flows
- Jospin, L.V., Laga, H., Boussaid, F., Buntine, W. and Bennamoun, M., 2022. Hands-on Bayesian neural networks—A tutorial for deep learning users.
- Izmailov, P., Vikram, S., Hoffman, M.D., Wilson, A.G.G.. (2021). What Are Bayesian Neural Network Posteriors Really Like?

Backup

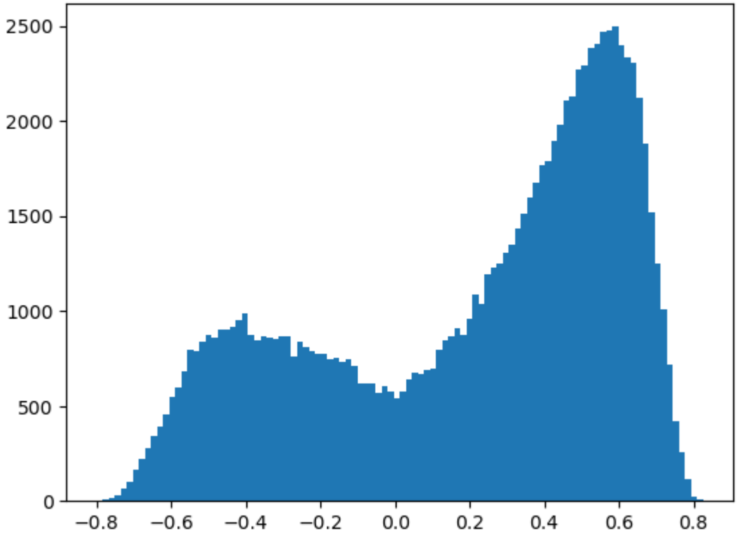
Example: softsign flow's effect on the Gaussian



N(0, 1)

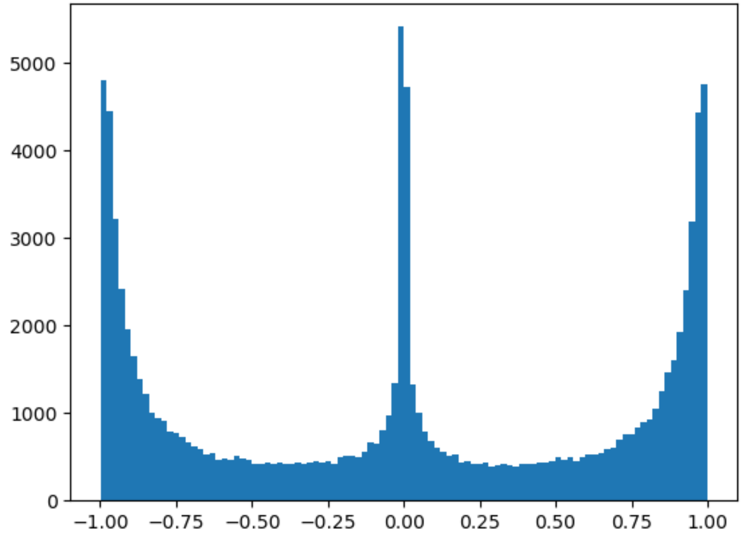


N(0, 0.5)

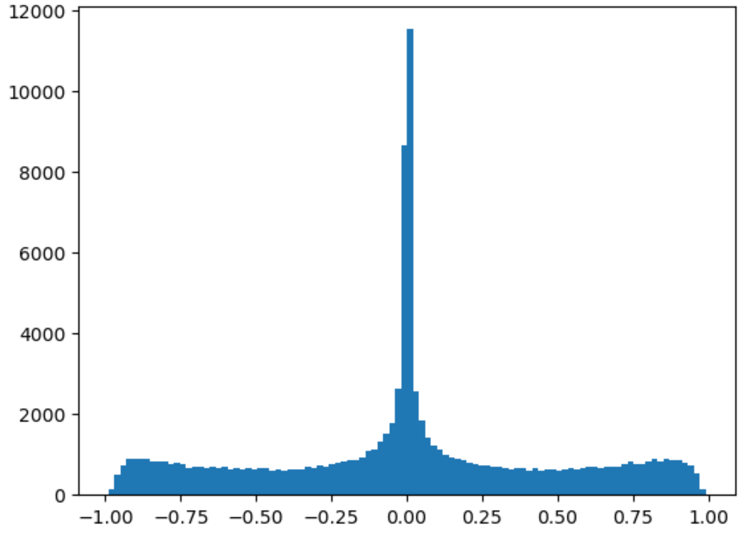


N(0.5, 1)

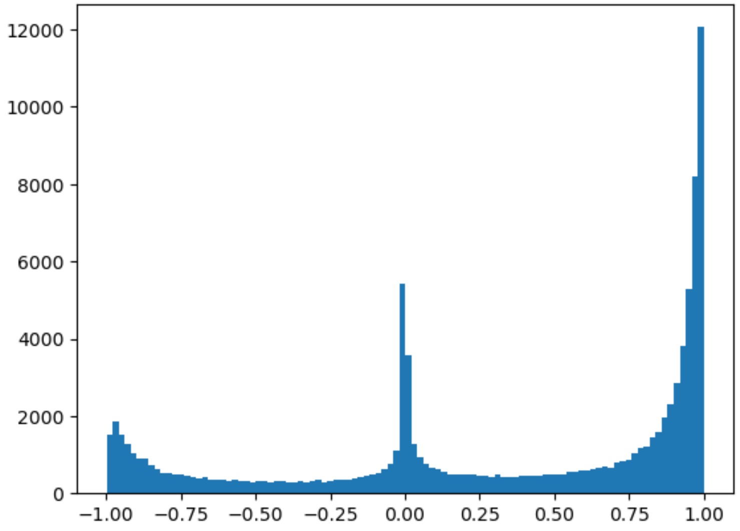
Example: softsign->scale(5)->power(3)



N(0, 1)



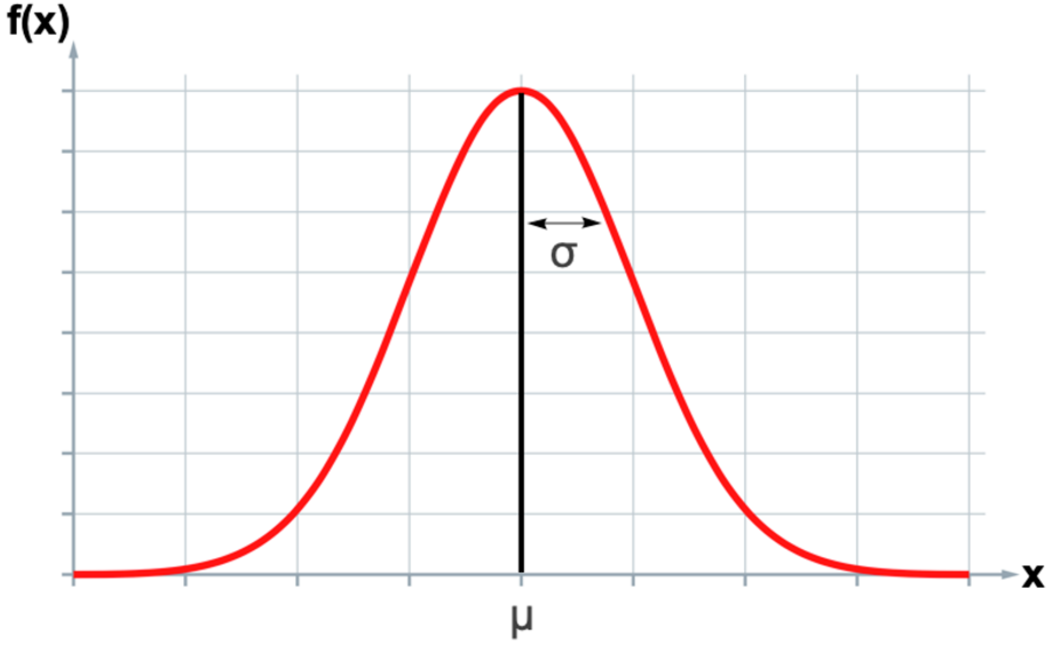
N(0, 0.5)



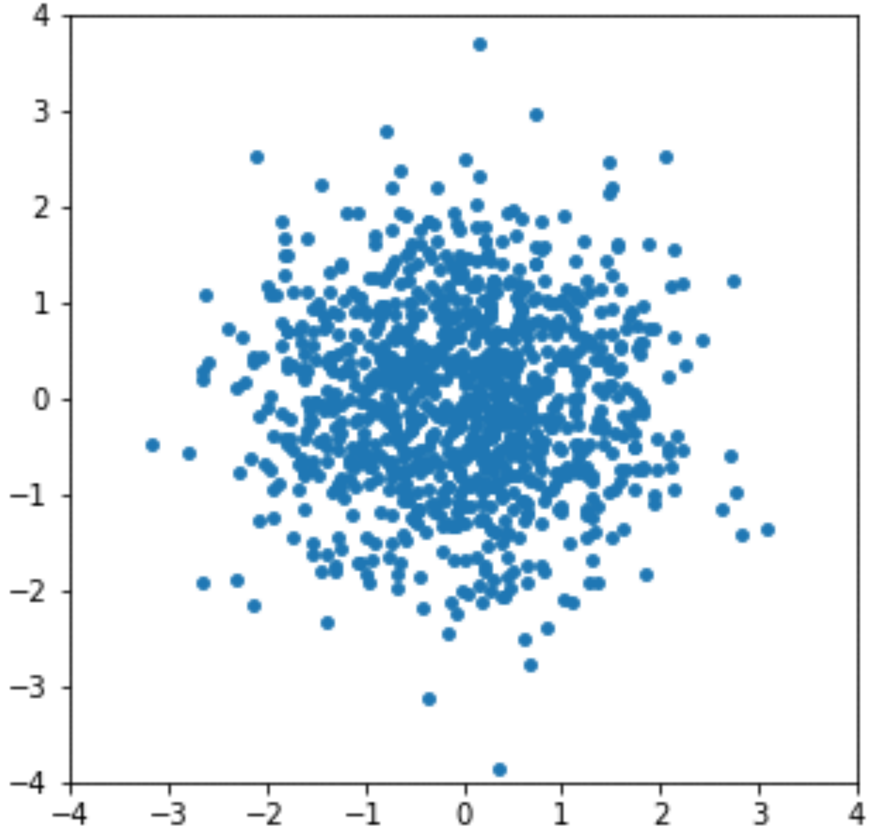
N(0.5, 1)

The "reparametrization trick" allows SGD methods to pass through stochastic nodes

$$w = \mu + \sigma * \epsilon, \qquad \epsilon \sim N(0, 1)$$



Visualization how samples from a Gaussian base distribution are transformed through 8 layers of Real NVP to match the target distribution:

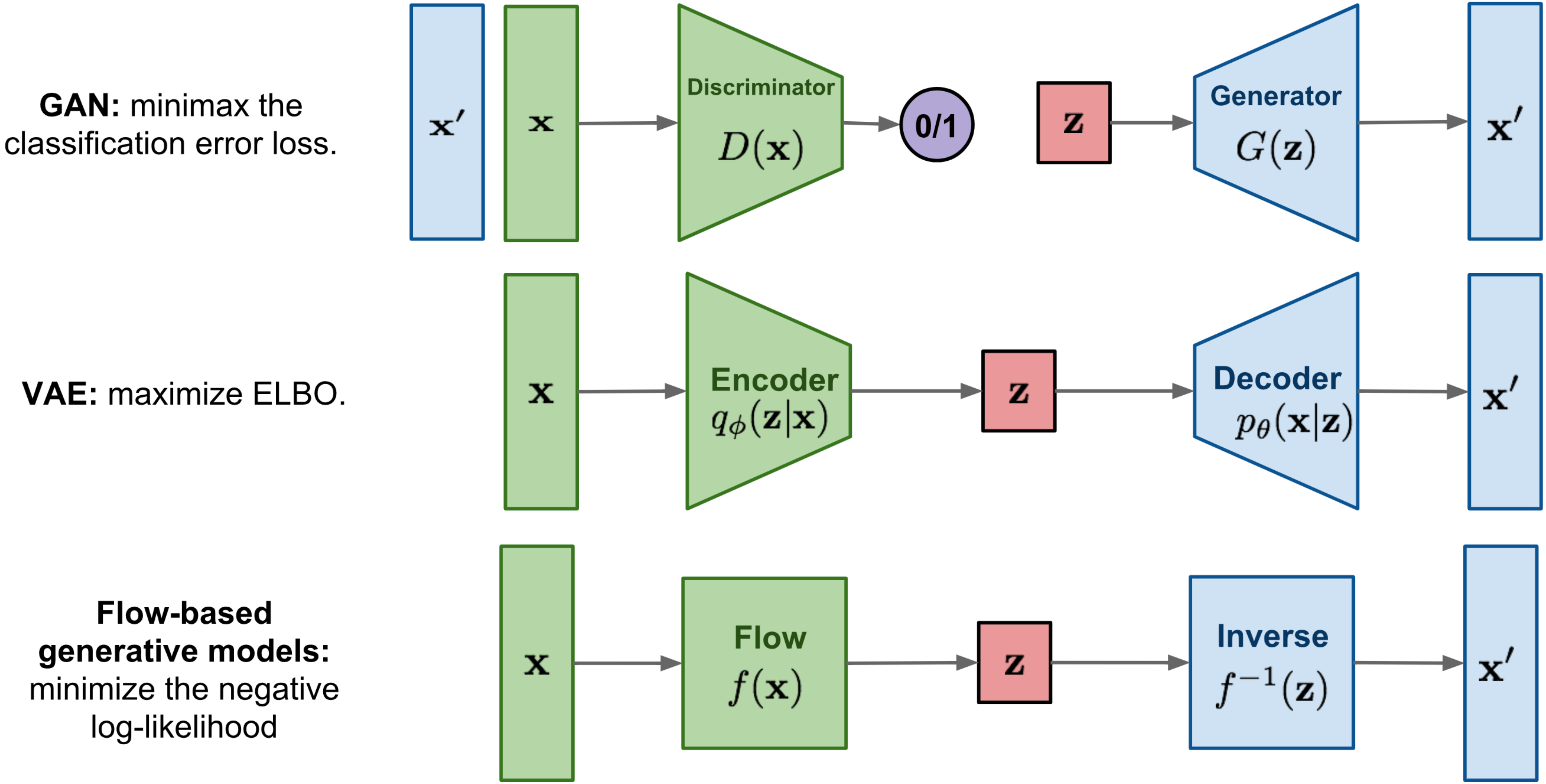


Posterior formula for a BNN – choosing the network's parameters

$$p(\boldsymbol{\theta}|D) = \frac{p(D_{\boldsymbol{y}}|D_{\boldsymbol{x}}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(D_{\boldsymbol{y}}|D_{\boldsymbol{x}}, \boldsymbol{\theta}')p(\boldsymbol{\theta}')d\boldsymbol{\theta}'} \propto p(D_{\boldsymbol{y}}|D_{\boldsymbol{x}}, \boldsymbol{\theta})p(\boldsymbol{\theta})$$

Predicting with a BNN – the marginal

$$p(\boldsymbol{y}|\boldsymbol{x}, D) = \int_{\boldsymbol{\theta}} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta'})p(\boldsymbol{\theta'}|D)d\boldsymbol{\theta'}.$$



GAN: minimax the classification error loss.
$\mathbf{x}'$
$\mathbf{x}$
Discriminator
$D(\mathbf{x})$
0/1
$\mathbf{z}$
Generator
$G(\mathbf{z})$
$\mathbf{x}'$
VAE: maximize ELBO.
$\mathbf{x}$
Encoder
$q_\phi(\mathbf{z}|\mathbf{x})$
$\mathbf{z}$
Decoder
$p_\theta(\mathbf{x}|\mathbf{z})$
$\mathbf{x}'$
Flow-based generative models: minimize the negative log-likelihood
$\mathbf{x}$
Flow
$f(\mathbf{x})$
$\mathbf{z}$
Inverse
$f^{-1}(\mathbf{z})$
$\mathbf{x}'$